A Level H2 Math

Correlation and Linear Regression Test 6

Q1

The consumer price index measures the average price changes in a fixed basket of consumption goods and services commonly purchased by resident households over time. It is commonly used as a measure of consumer price inflation. In the 2013 Singapore household expenditure survey, housing and food made up about half of the average monthly expenditure of an average household.

The table below shows the housing and food price index from 2005 to 2012, where 2005 is the base period, i.e. in 2005, the price index is 100. For example, the food price index of 104.6 in 2007 means that average food prices increased by 4.6% from 2005 to 2007.

Year	2005	2006	2007	2008	2009	2010	2011	2012
Housing Price Index, x	100	100.7	102.3	116.8	123.1	124.3		148
Food Price Index, y	100	101.6	104.6	112.6	115.2	116.8	120.3	123.1

(i) Show that the value of the missing housing price index for 2011 is 136 (nearest integer), given that the regression line of y on x is $y = 54.271 + 0.48363x$, correct to 5 significant figures. [2]

(ii) Draw the scatter diagram for these values, labelling the axes clearly. Comment on the suitability of the linear model. [3]

(iii) It is required to estimate the housing price index in 2016 where the food price index in 2016 is 134.6. Find the equation of an appropriate regression line for y and $\sqrt{x}$ and use it to find the required estimate. Explain why this estimate might not be reliable. [4]

(iv) Find the product moment correlation coefficient between y and $\sqrt{x}$. [1]

(v) To simplify recordings and calculations, it would be more convenient to tabulate $\frac{x}{100}$ and $\frac{y}{100}$ instead. Without any further calculations, explain if the product moment correlation coefficient between $\sqrt{\frac{x}{100}}$ and $\frac{y}{100}$ would differ from the value obtained in part **(iv)**. [1]

Q2

A study is done to find out the relationship between the age of women and the steroid levels in the blood plasma. Sample data collected from 10 females with ages ranging from 8 years old to 35 years old is as shown below.

Age (years) x	8	11	14	17	20	23	26	29	32	35
Steroid Level (mmol/litre) L	4.2	11.1	16.3	19.0	25.5	26.2	24.1	33.5	20.8	17.4

(i) Give a sketch of the scatter diagram for the data. Identify the outlier and suggest a reason, in the context of the question, why this data pair is an outlier. [3]

For the remaining part of the question, the outlier is to be removed from the calculation.

(ii) Comment on the suitability of each of the following models. Hence determine the best model for predicting the steroid level of a female based on her age.

Model A: $L = a + b\ln x$

Model B: $L = c + d(x-25)^2$

Model C: $L = e + f(x-25)^4$

where a, b, c, d, e and f are constants. [3]

(iii) Using the best model in **(ii)**, estimate the steroid level of a woman at age 40. Comment on the reliability of your estimate. [3]

(iv) It is known that body muscle mass and steroid level has a linear correlation. The muscle mass percentage m % of the 9 females were measured. An additional female, Jane, participated in the study. Jane has her muscle mass percentage and steroid level measured. The mean muscle mass percentage of the 10 females is now found to be 26.28 %. The equation of the least squares regression line of m on L for the 10 pairs of data is

$$m = 2.22 + 1.25L.$$

Calculate Jane's steroid level. [3]

Q3

A researcher investigates the relationship between the population of a particular species of bacteria in millions (b) and the surrounding temperature in °C (t). The researcher keeps records so that she can estimate the population of the bacteria at a certain temperature. Observations at different temperatures give the data as shown in the following table.

t	26.5	27.5	28.5	29.5	30.5	31.5	32.5
b	1.31	2.10	3.65	5.80	α	19.56	31.20

(i) Given that the regression line of b on t is $b = -129.368 + 4.75214t$, show that $\alpha = 12.12$, correct to 2 decimal places. [2]

(ii) Sketch a scatter diagram for the data. [1]

(iii) Explain which of $b = ct + d$ or $b = kt^3 + l$ is the more appropriate model for the relationship between b and t and find the equation of a suitable regression line for this model. [2]

(iv) Use the model you chose in part **(iii)** to estimate the population of the bacteria when the temperature is 33°C. Comment on the reliability of the estimate obtained. [2]

(v) It is given that the temperature T, in °F, is related to the temperature t, in °C, by the equation $T = 1.8t + 32$. Rewrite your equation from part **(iii)** so that it can be used to estimate the population of bacteria when the temperature is given in °F. [2]

Answers

Correlation and Linear Regression Test 6

Q1

(i) Let k be the missing housing price index for 2011.

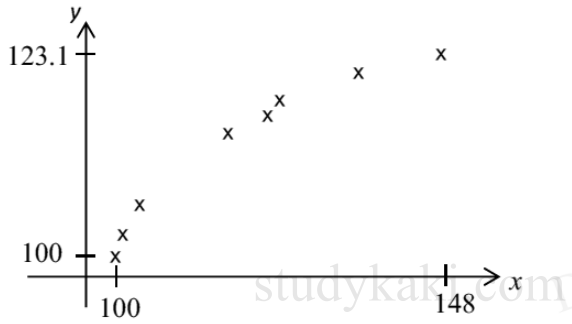
$\overline{y} = 111.775$ and $\overline{x} = \dfrac{815.2+k}{8}$

Since $\overline{y}$ and $\overline{x}$ lies on the regression line,

$$111.775 = 54.271 + 0.48363\left(\frac{815.2+k}{8}\right)$$

$k = 136$ (3 s.f.)(shown)

(ii)



From the scatter diagram, as x increases, y increases at a decreasing rate. Thus the linear model might not be the most appropriate model.

(iii) (Note that there is no clear independent variable.)
From GC, an appropriate regression line would be
$\sqrt{x} = 0.0896y + 0.860$ (3 s.f.)

When $y = 134.6$, from GC, $x = 167$ (3 s.f.).
The estimated housing price index in 2016 is 167.

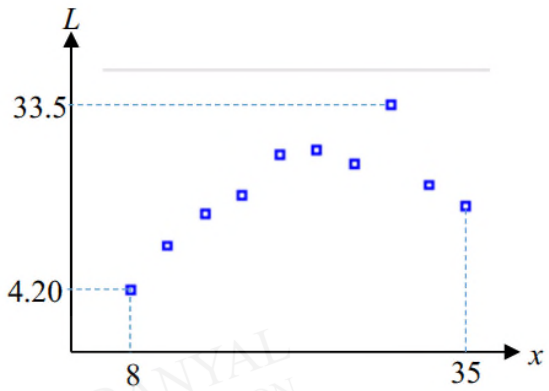
Since $y = 134.6$ <u>falls outside the data range of y</u>, the linear correlation between y and $\sqrt{x}$ might no longer hold and thus, the estimate is unreliable.

(iv) From GC, $r = 0.979$ (3 s.f.).

(v) The product moment correlation coefficient between $\sqrt{\dfrac{x}{100}}$ and $\dfrac{y}{100}$ **<u>does not differ</u>** from the value obtained in part **(iv)** as the r-value is **<u>independent of the scale of measurement.</u>**

Note that: $\sqrt{\dfrac{x}{100}} = \dfrac{\sqrt{x}}{10}$ means that the values of $\sqrt{x}$ undergo a scaling of 10 units and $\dfrac{y}{100}$ means that the values of y undergo a scaling of 100 units.

Q2

(i)  Outlier is $x = 29$, $L = 33.5$ because from age 23 onwards, there is a decreasing steroid level as age increases. However, at $x = 29$, the steroid level suddenly increases and this could be due to reasons such as illness/medication/pregnancy/intake of additional steroids by athlete/..etc (give any one of these reasons)	Students need to take note of what to indicate on scatter diagram: - Label of axes - Spacing and different in "height" between data points - Label min and max values The identification of outlier cannot be just circling of the point. Student must clearly states the x and L value of the outlier. Explanation of why (29, 33.5) is an outlier must be provided with a suggested possible contextual reason as well as an explanation of the kind of data trend that is resulted from this reason.
(ii) Model A is not suitable because as x increases, L is either only increasing (if $b > 0$) or only decreasing (if $b < 0$) which does not resemble the data trend of x and L whereby L increases but when it reaches about 23 years old, the steroid level decreases. Model B and C have similar trend as the given data set when $d < 0$ and $f < 0$ respectively and both are suitable models. But for Model C, the r-value of L and $(x-25)^4$ is -0.929 And for model B, the r-value of L and $(x-25)^2$ is -0.987 Since for model B, the r value is closer to -1, therefore model B is a better model.	Suitability of model must take into consideration the difference between the model and the data trend, using appropriate (sign) of b, d and f. As it is not possible to gauge the steepness of gradient base on the data points in the scatter diagram, thus the use of steepness to decide on whether model B or C is better is not accepted. Calculation of r must omit the outlier.

(iii) Least squares regression line equation is $L = 25.238 - 0.073667(x-25)^2$ $L = 25.2 - 0.0737(x-25)^2$ When $x = 40$, $L = 25.2 - 0.0737(40-25)^2 \approx 8.7$ The prediction is unreliable because $x = 40$ is outside the data range of 8 to 35 years old.	Calculation of equation of regression line must omit the outlier. Student must note that they cannot use the command $Y_1(40)$ directly to find L as the GC treat $(x-25)^2$ as X. Answer should be left to 1 decimal place, following that in the given table of L values.
(iv) Since $\overline{m} = 2.22 + 1.25\overline{L}$, $26.28 = 2.22 + 1.25\overline{L}$ $\Rightarrow \overline{L} = 19.248$ $\sum_{i=1}^{10} L_i = 192.48$ and since $\sum_{i=1}^{9} L_i = 164.6$, therefore Jane's steroid level is $27.88 \approx 27.9$	Students must realise that even if they are able to find the value of Jane's muscle mass t be 37.07, they cannot substitute this value into the equation to find Jane's steroid level. So even if the answer obtained is also 27.9, they are wrongly assuming that the data point lies on the regression line. Only $(\overline{L}, \overline{m})$ lies on the regression line.

Marker's comments

(i) The scatter diagram is quite well drawn but the labelling of axes, minimum and maximum values of x and L are often left out or wrongly labelled. Many students are mainly describing the data trend and the high L level of the point (29. 33.5) and did not give a contextual reason on why the data trend is as such. On the other hand, another group of students gave a very brief contextual reason but did not provide any elaborate on what this reason led to.

(ii) Many students did not explore the different possible sign of b, d and f. Most students wrongly assume that b, d and f are positive and rejected model B and C. One serious mistake that some students made is that they associate the power n in the expression $(x-25)^n$ to the number of turning points that the graph has, not realising that for all even integer n, there is only one turning point. Many students also did not read the question instruction to comment on the suitability of **each** model, they mainly compute values of r for all 3 models and conclude the one best model.

(iii) Many students left the estimated value of L to 3 s.f. instead of 1 decimal place. Some forgot to write down the equation of the regression line. Some wrongly write the equation as $L = 25.2 - 0.0737x^2$. Many forgot to omit the outlier in both part (iii) and in (ii) when finding r value.

Although most students are able to answer this part correctly, their answers are rather vague. Phrasing such as "it is an extrapolation" or "It is within data range" is not acceptable as it is unclear whether the student is referring to x or L within data range. Students must also remember to answer the question using the given term "not reliable" instead of "not accurate".

(iv) This part is generally well done but the notation of $\overline{L}$ is often not used, many just write it as L even if their subsequent workings show that they know that the value 19.248 is the mean steroid level.

Q3

9(i)

$b = -129.39 + 4.7529t$

From GC, $\bar{t} = 29.5$

$\bar{b} = -129.39 + 4.7529\bar{t}$

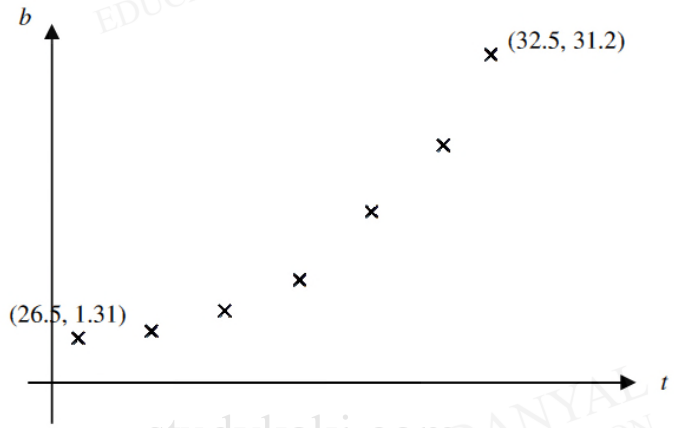
$\bar{b} = -129.39 + 4.7529(29.5)$

$= 10.82055$

$\frac{1.31 + 2.1 + 3.65 + 5.8 + \alpha + 19.56 + 31.2}{7} = 10.82055$

$\alpha = 12.124 = 12.12$ (2 dp)

(ii)



(iii)

From (ii), the scatter diagram shows that as t increases, b increases at an increasing rate which would not be the case if the data follows a linear model. Hence the model $b = kt^3 + l$ is a better model.

$b = -37.370 + 0.0018516t^3$

$= -37.4 + 0.00185t^3$ (3s.f.)

(iv)

When $t = 33$,

$b = -37.370 + 0.0018516(33)^3$

$= 29.171$

$= 29.2$ (3s.f.)

The population of the bacteria is 29.2 millions.

Since the estimate is obtained via extrapolation, the estimate is not reliable.

(v)

$b = -37.370 + 0.0018516\left(\frac{T-32}{1.8}\right)^3$

$= -37.370 + (3.1749 \times 10^{-4})(T-32)^3$

$= -37.4 + (3.17 \times 10^{-4})(T-32)^3$ (3 s.f.)